Supplementary Results

Canvas: versatile and scalable detection of copy number variants

Eric Roller[1,*,†], Sergii Ivakhno[2,†], Steve Lee[1,†], Thomas Royce[3,†] and Stephen Tanner[1,†]

[1]Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, [2]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL and [3]Ashion Analytics, 445 North 5th Street, Phoenix, AZ

*To whom correspondence should be addressed. †These authors contributed equally to the manuscript.

Wednesday, January 13, 2016

Contents

1. Test data generation

1.1 Simulation

Cancer genome

Phased SNV and indel variants from Platinum Genomes (PG) project [1] comprising an extended 17-member family have been used to split aligned reads into separate haplotypes for each chromosome. The phased haplotypes were created by using MERLIN linkage software [2] on informative SNV calls generated by GATK [3]. PG sample NA12882 sequenced to a 200x depth on a HiSeq 2000 system was used a source of aligned reads. The SNV variant selection went through a number of stringent criteria including:

1) Variant calls have no Mendelian inconsistencies within a pedigree.

2) Variants are called in all replicates for each sample (where available).

The splitting procedure generated separate bam files for each of the chromosomal haplotypes. The following parameters were provided as input to create simulated tumor samples from generated haplotype bams:

1) A bed file with ground truth copy numbers for each haplotype.

2) A purity value indicating the level of normal contamination. Normal contamination was re-created by mixing a “tumor” sample with a “normal” sample representing unphased NA12882 technical replicate (sequenced to a 40x depth on a HiSeq 2000 system).

3) Variant heterogeneity. The following features were altered to recreate specific types of tumor evolution by simulating different clonal bams and merging them afterwards:

1. Number of clones
2. Clonal evolutionary tree structure
3. Percentage of major clone
4. Percentage of private (heterogeneous) variants

Three ground truth bed files with genome-wide copy number values were used as input representing near-haploid (194 somatic CNVs), diploid (115 somatic CNVs) and pseudo-tetraploid (119 somatic CNVs) genomes. Copy numbers for these files were derived from previously sequenced and manually curated tumor samples to represent realistic scenarios of somatic genome rearrangements. Each of these files was then used to simulate three genomes of 80x depth with normal contamination of 20% (purity 80%), two clones and major clone abundances / percentage of private (heterogeneous) variants of 60% / 60%, 60% / 80% and 80% / 80% respectively. For simplicity, X and Y chromosomes were excluded from simulation.

The code for haplotype-based tumor simulation along with installation and usage instructions can be downloaded from HapMix repository at https://github.com/Illumina . The repository also contains ground truth files used in simulation.

Germline genome

Haplotype-based simulation of germline CNVs was carried out in a manner similar to cancer genome simulation, but without incorporating purity and heterogeneity parameters. Germline variants were sub-sampled from 1000 Genomes CNV reference panel [4] to an average of 3,400 variants per genome and a copy number distribution from zero to five. Target coverage was set to 40x. Regions with known CNVs in NA12882 were excluded from evaluation (as identified by Canvas and FREEC calls). We simulated a total of 6 genomes.

1.2 Read titration

Breast carcinoma cell lines HCC2218 / HCC1187 and corresponding normal lymphoblastoid cell lines HCC2218BL / HCC1187 BL were sequenced to 80x and 40x depth respectively on a HiSeq 2000 system. Reads were aligned using either BWA [5] or Isaac aligners [6] and titrated to simulate tumor purities of 80%, 60%, 40% and 20% (plus original samples, which were considered as 100% pure). Ground truth data was derived from manual inspection of coverage, minor allele frequencies (MAF) and paired-end read mappings. Difficult to interpret regions (i.e. telomeres) were excluded from evaluation: CN calls spanning those segments were not considered when deriving performance metrics.

1.3 Exome data

Whole-exome sequencing was carried out using Nextera Rapid Capture Exome Kit on a HiSeq 2500 system. HCC2218, HCC1187, HCC2218BL and HCC2218BL were sequenced at 120x, 90x, 100x and 130x respectively (depth for targeted regions). Only targeted regions specified in the manifest were evaluated.

1.4 Test data summary

In total, 37 samples were used in Canvas testing and methods comparison. This included 20 samples representing titrated HCC cell lines (10 samples aligned using either Isaac or BWA aligners), 9 samples from heterogeneity HapMix simulated tumors, 2 exome sequenced cell lines and 6 germline genomes. UCSC hg19 or Ensembl GRCh37 genome references have been used.

1.5 Simulation and real data comparison

To assess how realistically haplotype-mixing represents real data, we compared HCC2218 and HCC1187 calls made by Canvas on real data with simulated ones for the same samples. Genome-wide concordance between real and simulated data was 96.7% and 97.5% for HCC2218 and HCC1187 respectively. We then separately compared CNV calls for real and simulated data to the curated ground truth, results are shown in Table 1. As can be seen, simulation provided highly concordant to real data calls, corroborating its use in variant benchmarking.

Table 1. *Canvas somatic CNV metric comparison between HapMix simulated and real data.*

Data	HCC2218 accuracy	HCC1187 accuracy	HCC2218 precision	HCC1187 precision	HCC2218 recall	HCC1187 recall
Real sequencing data	98.54	99.07	98.25	98.96	95.79	98.26
HapMix simulation	99.06	98.79	99.08	99.48	97.70	97.64

1.6 Data access

Raw fastq files and aligned read data (using Isaac aligner [6]) from whole-genome sequencing of HCC2218 and HCC1187 can be accessed from BaseSpace at https://basespace.illumina.com/projects/5799796 . The corresponding data for Nextera Rapid Capture Exome preparation can be downloaded from https://basespace.illumina.com/s/DcPnOqHmtPNB . HCC2218 and HCC1187 cell lines were invented by Drs. Adi F. Gazdar and John D. Minna at the University of Texas Southwestern Medical Center. Rights in and to the HCC cell lines, progeny, and unmodified derivatives thereof belong to the Board of Regents of The University of Texas System. Illumina, Inc. has obtained permission from the Board of Regents of The University of Texas System through the University of Texas Southwestern Medical Center to use the HCC cell lines and publish the data and results herein displayed. For input into HapMix simulation, a technical replicate of NA12882 sequenced to 200x depth on a HiSeq 2000 system was used. Raw fastq files for the sample are available for download from European Nucleotide Archive under the accession number ERP001775.

1.7 Variant calls generation

All CNV callers were run according to usage instructions. If default ploidy values were required, they were set to 2 to enable consistent comparison between callers. We didn't aim to perform a comprehensive evaluation of somatic exome CNV callers as this was covered elsewhere [7]. However, comparative assessment of Nam *et al*. was used to select tools that showed superior performance: EXCAVATOR [8] and ADTEx [9]. EXCAVATOR v2.2 was run in "somatic" mode since paired normal samples are available. ADTEx v2.0 was run without the "--baf" option as it significantly decreased the accuracy. For the whole-genome somatic CNV identification, Control-FREEC v7.2 [10], TitanCNA v1.6.0 [11] and THetA v0.62 [12] were used.

2. Evaluation strategy

Concordance between ground truth and predicted copy number states was used to evaluate performance of CNV callers. We have resorted not to use non-variant predictions such as genome ploidy and sample contamination (purity) in our evaluation. First, not all methods provided such outputs. And second, interactions between ploidy, purity and heterogeneity are often non-linear and their effects on downstream CN assignment are difficult to fully interpret. For instance, minor differences in expected and predicted purities might not lead to any copy number differences and the tool that consistently underpredicts purity can still achieve 100% accuracy as far as variant calling and downstream annotation/interpretation are concerned.

For benchmarking we have developed a tool called EvaluateCNV that uses a per-base assessment of concordance to derive performance metrics. The EvaluateCNV tool is included with the Canvas source distribution on GitHub.

Briefly, EvaluateCNV accumulates a 2D matrix of base counts CN[TruthSetCN, CalledCN]. For instance, CN[2, 3] is the number of bases where the truth set has copy number 2, but the CNV caller assigned copy number 3. For each X, CN[X,X] is the number of bases where a correct copy number X was called. EvaluateCNV excludes any bases that are present in the set of excluded segments or are not covered by any interval in the truth set. For each CNV call, EvaluateCNV then finds any intervals in the truth set that it overlaps, measures the length of the overlap, and subtracts off any portion of the overlap interval which is covered by the excluded intervals. The remaining bases are added to the count CN[TruthSetCN, CalledCN]. We assume that any intervals not included in the CNV calls were assigned copy number 2.

Once the CN array has been populated, EvaluateCNV accumulates and reports various metrics as follows:

- Accuracy: Total bases for all cells CN[X, X] divided by total bases

- Precision: Total bases for all cells CN[Y, Y] for Y != 2, divided by total bases CN[X, Y] for Y != 2
- Recall: Total bases for all cells CN[X, X] for X != 2, divided by total bases CN[X, Y] for X != 2
- Gain precision: Total bases for all cells CN[Y, Y] for Y > 2, divided by total bases CN[X, Y] for Y > 2. (Similarly for loss precision)
- Gain recall: Total bases for all cells CN[X, X] for X > 2, divided by total bases CN[X, Y] for X > 2. (Similarly for loss precision)
- Directional accuracy, precision, and recall are defined similarly, except that instead of requiring the copy numbers X and Y be equal, we require either (a) X<2 and Y<2, (b) X=2 and Y=2, or (c) X>2 and Y>2. In other words, directional accuracy only requires that gains are called as gains, so truth set copy number 5 and CNV call copy number 6 is considered correct.

Centromere and telomere regions were excluded from assessment, as well as regions with known CNV for NA12882 in the case of haplo-simulated data. In the latter case, CNVs were called using Canvas in germline mode.

3. Results

3.1 Whole Genome Tumor

Table 2 shows average values for various performance metrics of different CNV callers derived using EvaluateCNV across all samples. A similar assessment for samples with low purity and polyclonality is shown in Table 3, while Figure 1 and 2 show distribution of these metrics. Among compared tools Canvas had the largest number of best performing metrics when considering all samples: three for Canvas versus two for TITAN and one for FREEC. Moreover, while the distance between Canvas and second best performing tool was 13.3% for metrics where Canvas performed best, it was only 2.7% for metrics where Canvas didn't show the best result.

Table 2. *Comparative assessment of somatic CNV callers.*

Metrics	Canvas	Theta	TITAN	FREEC	FREEC contamination
Accuracy	**86.81**	40.49	70.31	67.15	72.69
Precision	**77.09**	17.94	58.91	34.64	50.32
Recall	67.93	29.74	**68.51**	25.75	42.02
DirectionAccuracy	**90.50**	50.15	79.31	81.54	82.89
DirectionPrecision	90.71	32.42	75.33	91.58	**94.79**
DirectionRecall	81.81	51.84	**86.07**	57.02	72.69

Table 3. *Comparative assessment of somatic CNV callers on samples with mid-to-low purity (>=60%) and heterogeneity.*

Metrics	Canvas	Theta	TITAN	FREEC	FREEC contamination
Accuracy	**82.79**	39.57	71.85	64.70	63.87
Precision	**71.80**	14.80	63.43	33.83	40.02
Recall	65.60	23.28	**74.45**	19.94	26.65
DirectionAccuracy	**85.91**	48.33	77.50	72.93	70.42
DirectionPrecision	86.64	28.45	77.80	85.67	**88.93**
DirectionRecall	73.07	43.71	**85.48**	39.29	40.69

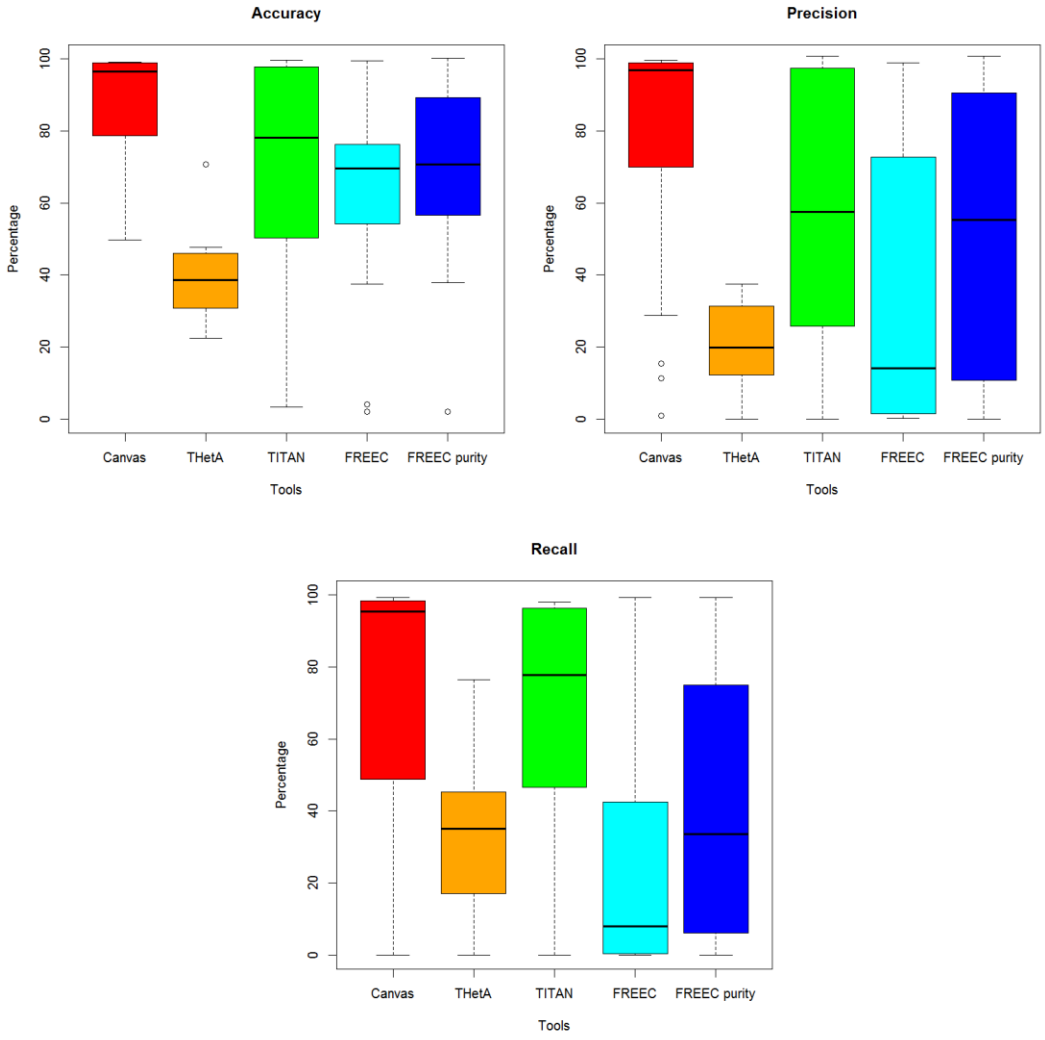

Figure 1. *Distribution of exact copy number performance metrics in different somatic CNV callers*

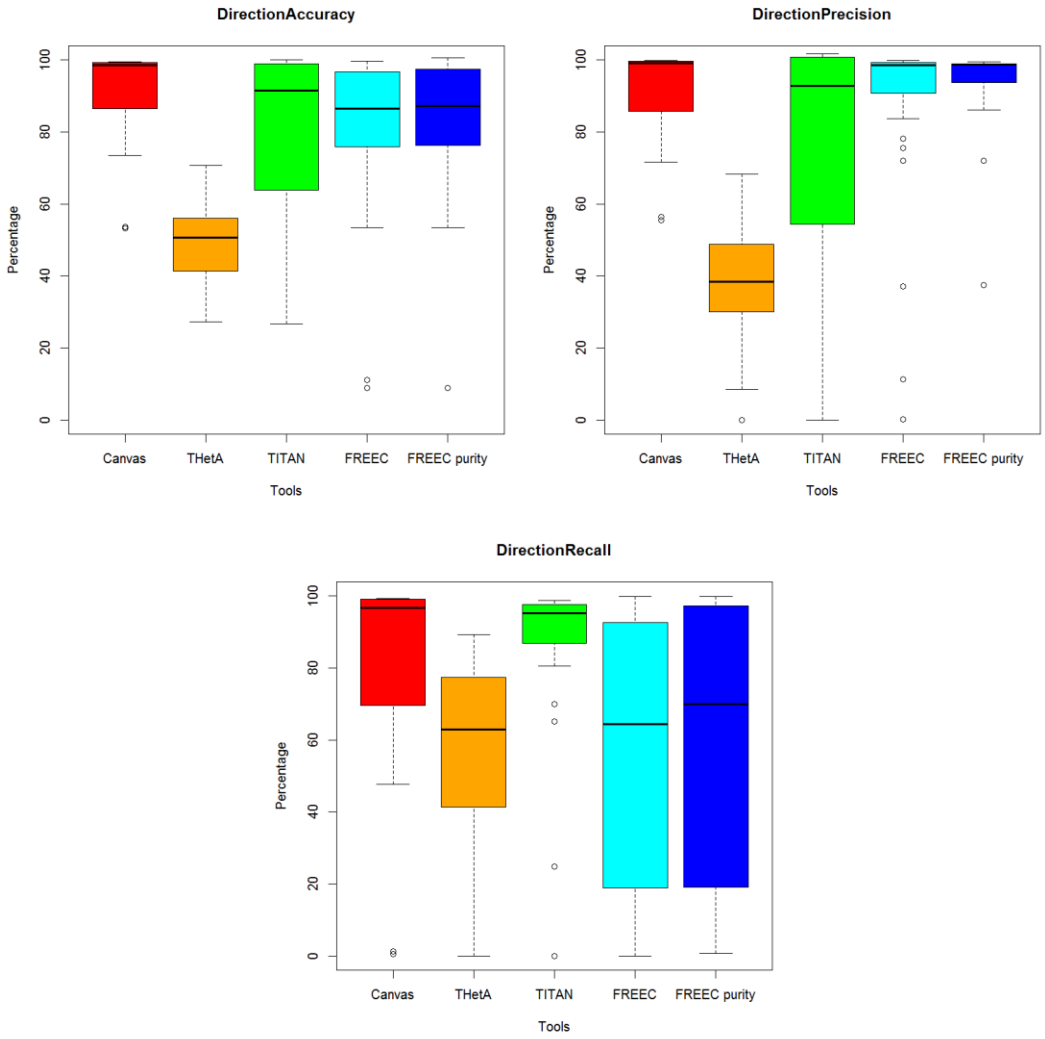

Figure 2. *Distribution of directional copy number performance metrics in different somatic CNV callers*

3.2 Whole Genome Germline

We have assessed Canvas re-sequencing workflow performance on HapMix simulation data and compared the results with those from FREEC. Canvas showed superior results on all metrics used, in particular it featured a much lower number of false negatives (Table 4).

Table 4. *Whole-exome comparative performance of germline CNV callers using haplotype simulation data.*

Metrics	Canvas	FREEC
Accuracy	93.08	63.86
DirectionAccuracy	94.03	64.95
Recall	92.91	63.86
DirectionRecall	94.12	64.95
Precision	98.42	97.86
DirectionPrecision	99.60	99.52

3.3 Whole Exome Tumor

Table 5 and 6 list comparative performance of Canvas, ADTEx and EXCAVATOR for HCC1187 and HCC2218 respectively. As can be seen Canvas consistently attained largest number best performing metrics across both samples.

Table 5. *Whole-exome comparative performance of somatic CNV callers on HCC1187 cell line. Best performing figures for each metric are highlighted in bold.*

Metrics	ADTEx	ADTEx (BAF)	EXCAVATOR	Canvas
Accuracy	97.42	0.79	54.96	**97.64**
DirectionAccuracy	99.60	45.10	82.35	98.97
Recall	94.80	1.56	3.67	**95.07**
DirectionRecall	**99.46**	96.37	62.28	97.93
Precision	95.00	0.74	5.89	**96.95**
DirectionPrecision	99.67	45.78	**99.92**	99.86

Table 6. *Whole-exome comparative performance of somatic CNV callers on HCC2218 cell line. Best performing figures for each metric are highlighted in bold.*

Metrics	ADTEx	ADTEx (BAF)	EXCAVATOR	Canvas
Accuracy	84.41	21.23	88.68	**90.58**
DirectionAccuracy	88.21	49.50	91.49	**94.73**
Recall	83.54	28.26	**89.77**	82.76
DirectionRecall	96.15	88.76	**99.11**	96.55
Precision	63.56	14.96	70.98	**74.74**
DirectionPrecision	73.16	46.98	78.3	**87.20**

3.4 Runtime Analysis

Runtime becomes an important factor when considering large-scale whole genome sequencing projects. We have analyzed runtime characteristics of Canvas germline and tumor-normal whole-genome workflows and compared their benchmarks with third party tools (Table 7). Canvas FFPE mode invokes fragment-based GC content normalization designed to pre-process formalin-fixed, paraffin-embedded tumor samples. Analysis was performed on a Linux CentOS 6.5 node with 32 cores, 126 gigabytes of RAM and EMC Isilon x410 storage system. Data for tumor-normal workflow included HCC2218/HCC1187 cancer-normal cell line pairs sequenced to an average of 80x and 40x respectively. Germline re-sequencing workflow included haplo-simulated samples described in *Simulation* section sampled to an average depth of 40x. All tools were run in a parallel mode, where such a feature was available, with utilization of all available CPUs.

Table 7. *Comparative runtime (minutes) of CNV calling tools.*

Workflow type	Canvas	Canvas FFPE	Theta	TITAN	FREEC
Tumor-normal	**40**	**72**	258	102	168
Re-sequencing	**35**	NA	NA	NA	138

References

[1] http://www.platinumgenomes.org

[2] Abecasis, G. *et al*. (2002) Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.,* **20**, 97-101.

[3] McKenna, A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* **20**, 1297-1303.

[4] Handsaker, R. *et al*. (2015) Large multiallelic copy number variations in humans. *Nat Genet.,* **47**, 296-303.

[5] Li, H. *et al*. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25**, 1754-1760.

[6] Raczy,C. *et al*. (2011) Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics,* **29**, 2041-2043.

[7] Nam, J. *et al*. (2015) Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief Bioinform.,* doi: 10.1093/bib/bbv055.

[8] Magi, A. *et al*. (2013) EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol*, doi:10.1186/gb-2013-14-10-r120.

[9] Amarasinghe, KC. *et al*. (2014) Inferring copy number and genotype in tumour exome data. *BMC Genomics*, **15**, 732.

[10] Boeva, V. *et al*. (2011) Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data. *Bioinformatics,* **28**, 423-425.

[11] Ha, G. *et al*. (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881-1893.

[12] Oesper, L. *et al*. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology,* doi:10.1186/gb-2013-14-7-r80**.**